维吾尔文论坛中基于术语选择和 Rocchio 分类器的文本过滤方法 *

如先姑力・阿布都热西提 [1]，亚森・艾则孜 [1†]，艾山・吾买尔 [2a]，阿力木江・艾沙 [2b]

(1. 新疆警察学院 信息安全工程系，乌鲁木齐 830013; 2. 新疆大学 a. 信息科学与工程学院; b. 网络中心，乌鲁木齐 830046)

摘 要：针对维吾尔文网页论坛中的文本过滤问题，提出一种基于术语选择和 Rocchio 分类器的文本过滤方法。首先，将论坛文本进行预处理以删除无用词，并基于 N-gram 统计模型进行词干(术语)提取；然后，提出一种均衡考虑相关性和冗余性的均衡型互信息术语选择方法(BMITS)，对初始术语集合进行降维，获得精简术语集；最后，将文本特征术语作为输入，通过 Rocchio 分类器进行分类，以此过滤掉论坛中的不良文本。在相关数据集上的实验结果表明，提出的方法能够准确地识别出不良类型文本，具有有效性。

关键词：维吾尔文论坛；文本过滤；N-gram 统计模型；术语选择；Rocchio 分类器

中图分类号：TP391　　doi: 10.3969/j.issn.1001-3695.2017.10.0940

Text filtering method based on term selection and Rocchio classifier in uyghur forum

Ruxianguli·abudurexiti[1], Yasen·aizezi[1†], Aishan·wumaier[2a], Alimu·aisha[2b]

(1. *Dept. of Information Security Engineering Xinjiang Police College*, *Urumqi* 830013, *China*; 2. *a. School of Information & Engineering*, *b. Network Centry*, *Xinjiang University*, *Urumqi* 830046, *China*)

Abstract: For the issues that the text filtering in Uyghur web forum, this paper proposed a text filtering method based on term selection and Rocchio classifier. Firstly, it preprocessed the forum text to remove useless words and extract stemming (term) based on the N-gram statistical model. Then, it proposed a balanced mutual information term selection method (BMITS) , which considered the correlation and redundancy of equilibrium, used to reduce the dimension of initial term set and obtain the reduced term set. Finally, it made the text feature terms as input, and used Rocchio classifier to filter out the bad text. The experimental results show that the proposed method can accurately identify the bad type text, which is effective.

Key Words: uyghur forum; text filtering; N-gram statistical model; term selection; Rocchio classifier

0 引言

随着互联网的高速发展，网页论坛也爆发式增加。论坛方便了网民的信息交流，也提高了工作学习效率。然而，由于网页论坛是开放式的，也存在一些负面影响，如迷信、反动、暴力色情等非法信息的传播，不利于社会的稳定和人民身心健康[1,2]。因此，对网页论坛中一些非法文本进行过滤具有重要的意义。

近些年，随着国家对新疆地区发展的大力支持，网络化建设也得到快速发展，产生了很多以维吾尔文进行书写的 Web 论坛。有些疆独分子通过维吾尔文论坛传播着各种不良信息，为此，对维文网页论坛中的不良文本进行过滤，对新疆的长治久安具有促进作用[3]。

为了实现维吾尔文网页论坛中的文本过滤，主要是对这些文本进行有效分类，然后将分类为不良类的文本进行删除[4]。对于维吾尔文文本分类的研究，近些年学者提出了一些方法。例如文献[5]提出了一种基于组合统计量(Dme)的维吾尔文文本分类方法，该 Dme 包含了互信息、t-测试和熵值，以此来进行词干提取和降维，并采用 K 近邻算法(k nearest neighbor, k-NN)作为文本分类器。文献[6]提出了一种基于词频-逆文本频率(term frequency-inverse document frequency, TF-IDF)和支持向量机(support vector machine, SVM)的维吾尔文情感分析方法，通过 TF-IDF 获得区分性关键词。文献[7]提出了一种基于 N-gram 模型和曼哈顿(Manhattan)距离的维吾尔文文本分类技术，其采用了 4 元模型，并在 Manhattan 距离中融入了骰子测量。然而这些方法都不能很好地对特征进行降维，导致文本分类精度不高且计算量较大。

为此，本文提出一种基于术语选择和 Rocchio 分类器的文

基金项目：国家自然科学基金资助项目(61762086)；国家社会科学基金资助项目(13CFX055)；新疆维吾尔自治区高校科研计划重点项目(XJEDU2017M046)

作者简介：如先姑力・阿布都热西提（1976-），女，新疆喀什人，副教授，硕士，主要研究方向为信息安全等；亚森・艾则孜（1975-），男（通信作者），新疆库车人，国家电子数据司法鉴定员，教授，硕士，主要研究方向为数字取证、自然语言处理等（yasenaizezi@126.com）；艾山・吾买尔（1981-），男，新疆库车人，副教授，博士，主要研究方向为自然语言处理；阿力木・艾沙（1973-），男，新疆喀什人，副教授，博士，主要研究方向为人工智能.

本过滤方法，并通过相关对比实验证明了该方法的有效性。提出方法的主要研究内容如下：

a)通过 N-gram 统计模型进行词干(术语)提取，并通过实验确认，当 N=4 时最适合维吾尔文文本特性。

b)为了解决传统基于互信息术语选择方法(mutual information term selection, MITS)的缺陷，提出一种均衡考虑相关性和冗余性的均衡型 MITS (balanced MITS, BMITS)，从初始术语集合中选择出具有高区别性的术语子集。

c)选择了在效率和泛化能力方面都较为优越的 Rocchio 分类器对文本进行分类，过滤掉不良文本。

1　维吾尔文的文本分类描述

1.1　维吾尔语的语言结构

维吾尔语是一种高度黏着性语言，其单词由 32 个字母组成，每种字母有 4 种不同的形式，致使其时态和形态变化比英语更丰富。维吾尔语中，通过在单词的结尾添加不同的词缀来实现语法功能[8]。即很多词语是由同一词根演变而来的，且这些单词的词义相差不大。由于这些特征，导致维吾尔语文本的原始特征维数大、文本表示稀疏等问题[9]，与传统中文或英文的文本分类方法相差很大。

维吾尔语的动词和一部分名词是由词根中形成的。词汇具有固定模式，通过在词的前后添加前缀和后缀可以表示它的数、性和时态。表 1 展示了可能添加到单词“شائىر”（诗人）中的不同词缀及其含义。其中，下画线上的字母为词缀。

表 1　词干“شائىر”（诗人）上的添加不同词缀形成的单词

维吾尔语单词	词义	维吾尔语单词	词义
شائىر	诗人	شائىردا	在诗人
شائىرە	诗人（女）	شائىرەدە	在诗人（女）
شائىرنىڭ	诗人的	شائىردەك	像个诗人
شائىرلار	诗人们	شائىرىم	我的诗人
شائىرەلەر	诗人们（女）	شائىرىڭ	你的诗人
شائىرلارنىڭ	诗人们的	شائىرەسى	他的诗人

1.2　维吾尔语的文本分类过程

已有大量的研究人员对汉语和英语文本进行分类研究，但很少有人对维吾尔文进行文本分类。在这里将对维吾尔文文本分类的三个主要步骤进行描述，分别为词干提取、特征降维和文本分类。

词干提取，其是从一个词中移除所有词缀来获得词根的过程，以此在文本信息获取任务中提高性能，特别是在类似于维吾尔语之类的高度黏着性语言中。在中文和英语文本分类研究中，词干提取大多采用去除后缀和停留词的方法。基于词根的词干提取技术是使用形态学分析方法对给定维吾尔语单词进行词根提取的操作。例如，文献[10]尝试通过将单词与所有可能的模式以及所有可能附加的词缀进行匹配，从而找到单词的词根，但是该算法不能删除任何前缀或后缀。文献[11]在形态分析系统中使用了不同的算法来找到词根和模式。其首先删除最长前缀，然后通过检查单词的前五个字母来提取词根。然而该算法基于一个假设，即词根一定会出现在单词的前五个字母中。在统计提取法中，常用的为 N-gram 模型[12]，其根据字符串相似性度量对相关单词进行分组。N-gram 模型是从单词中提取一组 N 个连续字符，相似的词将具有很高的 N-gram 比例。

特征降维，其是用来降低所提取的词干集合的维度，获得精简特征集。在移除停止词并提取词干后，将每个文本由一个具有 N 个权重项的矢量表示，称做文本表示的词包方法。在这一过程中，将会忽略文本的结构和词序，其特征向量表示文本中观察到的单词。训练集中的超级矢量 $W(w_1,\ldots,w_d)$ 由训练集中所有样本词干（也叫做术语）构成。通常，在文本分类中会有大量术语，因此，需要对术语空间进行降维。在英语文本分类中，通常通过一些术语评估函数来对术语集进行降维，选择出重要术语。这些函数有文本频率、互信息增益、χ^2 统计量、NGL 系数和 GSS 系数等[13]。

文本分类，其是根据输入的文本特征，对文本进行分类。目前常用的是通过对已经手动分类过的文本进行归纳学习，从而训练分类器。构建分类器具有两种不同的方法，即参数方法和非参数方法。参数方法中，训练数据用于估计概率分布的参数，如概率朴素贝叶斯分类器。非参数方法中，又可以进一步分为两类：基于样本的非参数方法，即将被分类的文本与训练集文本进行比较，将文本分类到与此文本相似度最高的训练文本类中，如 k-近邻分类器；基于描述的非参数方法，其首先将分类描述（或线性分类器）用一个术语权重的矢量表示，这一矢量通过对训练集文本预分类得到；然后将描述用做训练数据，并与待分类文本进行比较来进行分类，如 Rocchio 分类器。

2　提出的维吾尔语文本分类模型

本文目的是应用机器学习方法对维吾尔文网页论坛中的文本进行分类过滤。所提出的模型主要包含文本的预处理(词干提取)、术语选择和文本分类三个阶段。图 1 展示了所提出的维吾尔语文本分类模型。

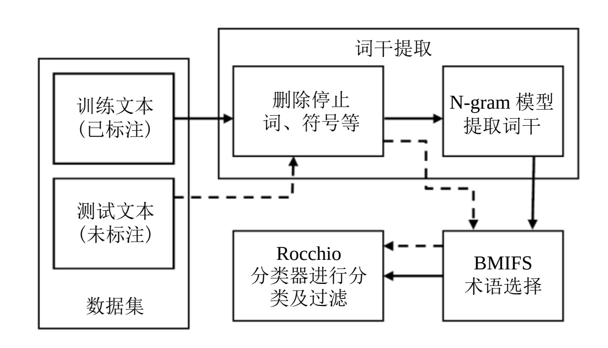


图 1　所提出维吾尔语文本分类系统的模型

文本预处理过程即词干提取过程，包含移除停止词和具有共同词根的词。之后，将构建一个超级矢量，再使用特征选择技术对超级矢量进行降维，并将文本以术语权重矢量的形式表

示。最后，构建分类器并评估其性能。

2.1 基于 N-gram 统计模型的词干提取

对于词干提取，通常有基于词根的方法和基于统计的方法，相比而言，基于统计的方法更适合维吾尔语文本分类任务。本文采用了 N-gram 统计模型来提取维吾尔语词干。采用的 N-gram 为字母级别，将所有连续的 N 个字母序列作为一个单元，称为一个 gram。

N-gram 模型中，其设定一个字母单元 l_i 在文本中出现的概率只与前 N-1 个字母相关。因此，字母序列 $L=l_1l_2l_3,...,l_N$ 出现的概率表示为

$$P(L)=P(l_1l_2l_3,...,l_N)=\prod_{i=1}^{N}P(l_i\,|\,l_{i-N+1},...,l_{i-1}) \tag{1}$$

在维吾尔语中，由于字母相互结合的概率很高，所以较短的 N 不能很好地表现单词属性，而 N=3,4 等较长时具有较强的代表性。

在本文基于 N-gram 统计模型的词干提取方法中，首先移除了单词中最常见的前缀和后缀，也包含外国语、数字、停止词等；然后通过 N-gram 模型计算两个词的相似性，以此来提取词干。基于 N-gram 统计模型的词干提取算法如算法 1 所示。

算法 1 基于 N-gram 统计模型的词干提取算法

For 文本中的每个词

If 非维吾尔语词汇 **Then** 该词是无用词；

If 包含数字 **Then** 该词是无用词；

If 单词长度<3 **Then** 该词是无用词；

移除附加符号，并标准化词汇；

If 该词是停止词 **Then** 该词是无用词；

移除前缀和后缀；

If 该词是停止词 **Then** 该词是无用词；

利用 N-gram 统计模型计算单词间相似性获得词干；

End For

首先算法确保单词是一个维吾尔语词，并认为长度少于三个字母的词在文章中是不重要的；接着会移除各种附加符号，这些符号在字母的上面或下面用于正字法，作为词法的标志；之后应用词标准化方法，将一些字母的不同写法（扩展区）统一为相同的形式，如将ﺱ， ﺲ， ﺳ， ﺴ 统一为س；将ﻫ，ﻪ 统一为 ﻬ等。

词形标准化后，算法会检查单词是否在一个停止词表中。停止词表由 165 个单词组成。消除停止词后，算法移除一组前缀(بى، نا، مر، ئاي، سۈپ، قاپ، تاپ 等)。移除后，算法会检查单词长度是否小于 3 个字母，如果小于 3 个，说明前缀是单词的一个主要部分，因此移除的前缀会恢复到单词中。接着将后缀(نىڭ، دا، تا، دىن، تىن، داش، چى، خان، غان، گەن، لار، لەر، ى، مى 等)递归地从词尾移除。首先从最长的后缀开始，再移除较短的。当词的前缀和后缀都移除之后，算法还会检查该词是否属于停止词表中的词汇，这是因为一些停止词也会附加前缀和后缀。

最后，利用 N-gram 统计模型计算单词间的相似性获得最终词干。对语料库中的所有术语对，计算其相似性度量。具有高于预定义相似性阈值的术语被聚类，并仅用其中一个术语来表示。

下面的例子描述了基于 N-gram 模型(N=2)，计算两个词 سىياسەت（政治）和 سىياسىنىڭ （政治的）的相似性。

1.سىياسەت⇒ سى، يا، سە، ت 。（首先将词分解为两字母组合模型）

2.分解成的两字母组合 ⇒ سى، يا، سە، ت。

3. سىياسىنىڭ ⇒ سى، يا، سى، نى، ڭ 。

4.分解成的两字母组合 ⇒ سى، يا، سى، نى، ڭ 。

那么，相似性为：$S=\dfrac{2C}{A+B}=\dfrac{2\times3}{4+3}=0.8571$。其中：$A$ 和 B 分别表示第一个词和第二个词中不同的两字母组合数量；C 表示两个词共同的两字母组合数量。将相似性大于阈值 T_S 的两个词归为一个词干。

2.2 基于 BMITS 的术语选择

2.2.1 传统术语选择方法

为了提高分类器的性能，需要对输入文本的术语集进行降维。术语选择技术用于从初始术语集中选择出最能表达文本意思的术语子集。通常使用术语评估函数 f_{TEF} 对初始集合中每个术语进行评分，选择出评分较高的术语。

在已有研究中，常用的特征降维技术有互信息、χ^2 统计量、NGL 系数以及 GSS 系数等方法，这些方法的表达式如下：

互信息增益 MI 为

$$MI(t_k,c_i)=\sum_{C\in\{c_i,\bar{c}_i\}}\sum_{t\in\{t_k,\bar{t}_k\}}P(t,c)\log\frac{P(t,c)}{P(t)P(c)} \tag{2}$$

χ^2 统计量 CHI 为

$$CHI(t_k,c_i)=\frac{|T_r|[P(t_k,c_i)P(\bar{t}_k,\bar{c}_i)-P(t_k,\bar{c}_i)P(\bar{t}_k,c_i)]^2}{P(t_k)P(\bar{t}_k)P(c_i)P(\bar{c}_i)} \tag{3}$$

NGL 系数 NGL 为

$$NGL(t_k,c_i)=\frac{\sqrt{|T_r|}\bullet[P(t_k,c_i)P(\bar{t}_k,\bar{c}_i)-P(t_k,\bar{c}_i)P(\bar{t}_k,c_i)]^2}{\sqrt{P(t_k)P(\bar{t}_k)P(c_i)P(\bar{c}_i)}} \tag{4}$$

GSS 系数 GSS 为

$$GSS(t_k,c_i)=P(t_k,c_i)P(\bar{t}_k,\bar{c}_i)-P(t_k,\bar{c}_i)P(\bar{t}_k,c_i) \tag{5}$$

其中：$P(\bar{t}_k,c_i)$ 表示为对一个文本 x，术语 t_k 不在 x 中，但是 x 属于 c_i 类的概率。

2.2.2 互信息术语选择方法(MITS)

互信息(MI)可表示一个随机变量中包含另一个变量信息的程度，是统计相关性的测度。其输出为一个非负值，其中零表示两个变量是统计独立的。

基于互信息理论来选择文本术语的方法称为互信息术语选择(MITS) [14]。MITS 中，其通过计算 $I(c;t_k,S)$ 来选择术语。$I(c;t_k,S)$ 表示在所选择的术语集 S 中，增加术语 t_k 后形成的新术语集与文本类别 c 之间的互信息，反映了术语 t_k 对文本分类的贡献程度。$I(c;t_k,S)$ 是通过计算术语 t_k 与类别 c 之间的互信

息 $I(c;t_k)$，然后计算术语 t_k 与先前所选术语 t_s 之间的互信息 $I(t_k;t_s)$ 之和，并将其乘以一个比例系数 β 对 $I(c;t_k)$ 进行校正来获得。表达式如下：

$$I(c;t_k,S)=I(c;t_k)-\beta\sum_{t_s\in S}I(t_k;t_s) \tag{6}$$

$I(c;t_k,S)$ 等式右边的第一项为候选术语与文本类别之间的互信息，表示相关性；第二项为求和候选术语与已选术语之间的互信息，表示冗余性。β 值表示为在进行术语选择时考虑输入术语之间冗余度的权重，其决定着在选择术语时，两个方面(即术语与文本类之间的 MI 和术语与术语之间的 MI)的重要性比重。

2.2.3 提出的 BMITS 方法

在传统 MITS 基础上，学者提出了几种改进型的 MITS 算法，如 MITS-U 算法等。这些方法大多是对式(6)中 $I(c;t_k,S)$ 中第二项进行了改进。然而这些方法存在一些限制。例如，$I(c;t_k,S)$ 中的相关性和冗余性通过一个参数 β 来进行权衡。如果 β 太小，则算法偏重候选术语与文本类之间的 MI，根据单个候选术语和文本类之间的 MI 依次选择术语；如果选择的 β 太大，则算法偏重考虑候选术语之间的 MI。为此 β 的选择较为困难，且目前也没有选择参数 β 的合适方法。

为了解决上述问题，本文提出一种均衡考虑相关性和冗余性的均衡型 MITS 算法(BMITS)，在第二项中引入了候选术语与文本类之间的互信息，且不再需要人为设置一个额外的参数，即利用 $1/|S|$ 代替 β。BMITS 从一个初始术语集中选择出具有最大化 $I(c;t_k)$ 并最小化冗余的术语，表达式如下：

$$G_{MI}=\arg\max_{t_i\in F}\left(I(c;t_k)-\frac{1}{|S|}\sum_{t_s\in S}MR\right) \tag{7}$$

其中：$|S|$ 为已选择术语的数量；MR 表示在已选术语集 S 中；术语 t_k 对于术语 t_s 的相对最小冗余，定义如下：

$$MR=\frac{I(t_k;t_s)}{I(c;t_k)} \tag{8}$$

当 $I(c;t_k)$=0 时，术语 f_i 可被丢弃。如果对于文本类 c，t_k 和 t_s 之间高度相关，那么 t_k 也将是冗余的。为此，需要设定一个阈值 $Th=0$ 来与 G_{MI} 进行比较。如果 $G_{MI}\le 0$，则当前术语 t_k 对于文本类 C 是不重要的，其将会降低所选择的术语集 S 与文本类 C 之间的 MI，并将其从 S 中删除；如果 G_{MI} >0，则将术语 t_k 作为候选术语。BMITS 选择术语的过程如算法 2 所示。

算法 2： BMITS 术语选择

输入：初始术语集 $T=\{t_k,k=1,\cdots,n\}$

输出：选择的术语集 S

开始

1. 初始化 $S=\phi$；
2. 为每个术语计算 $I(c;t_k)$；
3. 设置 $n_t=n$，根据下式选择术语 t_k：

   $\arg\max_{t_i}(I(c;t_k)),t=1,\cdots,n_t$；

   设置 $F\leftarrow F\setminus\{t_k\}$；$S\leftarrow S\cup\{t_k\}$；$n_t=n_t-1$；
4. **While** $F\ne\phi$ **do**

   计算互信息增益 G_{MI}，找到 $t_k,t\in\{1,2,\cdots,n_t\}$；

   设置 $n_t=n_t-1$； $F\leftarrow F\setminus\{t_k\}$；

   If ($G_{MI}>0$) **then**

   $S\leftarrow S\cup\{t_k\}$；

   End if

   End while
5. 根据 S 中每个术语的 G_{MI} 对术语进行排序并进行加权；
6. 返回 S。

2.3 基于 Rocchio 分类器的文本分类

Rocchio 分类器是一种典型的应用于文本信息过滤领域的分类器[15]。其会为每个类别 c_i 建立一个原型矢量，文本矢量 x 通过计算与每个原型矢量间的距离进行分类。类别 c_i 的原型矢量是根据属于类别 c_i 的所有文本矢量加权平均得到的。这意味着，与 k-NN 分类器相比，Rocchio 分类器具有更快的学习速度。

对于类别 $c_i(w_{i1},w_{i2},...,w_{in})$，其原型矢量可以根据下式计算得到

$$w_{ik}=\beta\bullet\sum_{d_j\in POS_i}\frac{w_{jk}}{|POS_i|}-\gamma\bullet\sum_{d_j\in NEG_i}\frac{w_{jk}}{|NEG_i|} \tag{9}$$

其中：w_{jk} 为术语 t_k 在文本 d_j 中的权重；POS_i 为属于第 c_i 类的文本集合（阳性样本）；NEG_i 为不属于第 c_i 类的文本集合（阴性样本）；β 和 γ 为控制参数，用来设置阳性样本和阴性样本的相对重要性。如果 β 设为 1 而 γ 设为 0，则类别 c_i 描述为其阳性训练样本的重心。Rocchio 分类器是根据阳性样本的聚集程度和阴性样本的疏远程度来进行分类的。阴性样本的作用通常是逆增强，这一效果通过设置较大的 β 值和较小的 γ 值得以体现。根据相关研究，可以设置 $\beta=1.6,\gamma=0.4$ [16]。

对于输入的未知类别样本，Rocchio 分类器通过比较输入样本 x 与每类原型矢量 w_{ik} 的最小距离来对样本进行分类。其中这个距离 $d()$ 通常为欧几里得距离。Rocchio 分类器的判决表示如下：

$$c_i^*=\arg\min_{c_i\in C}d(w_{ik},x) \tag{10}$$

Rocchio 算法通过引入一些拓展实例来解决 k-NN 算法的问题。即通过一个广义的实例代替整个训练样本集，这一广义实例是通过总结实例样本分布得到的。当新的实例加入进来时，对其分类只需要计算新实例与广义实例之间的欧氏距离即可。其时间复杂度为 $O(LM)$，其中 L 表示广义实例的数量，M 表示每个文本矢量中的术语数量。此外，根据每类中实例的分布，Rocchio 算法还可以解决噪声问题。例如，如果一个术语在某一类样本中频繁出现，就会同等反映在该类别的广义实例上，这个术语相对应的权值就会较高；另一方面，如果某一术语主要出现在其他类别的实例中，那么广义实例中这一术语的权值就会趋于 0。因此，Rocchio 分类器可以在一定程度上提取某些相关术语。

3　实验与分析

3.1　维吾尔文论坛文本集

本文在 Matlab2014 软件上实现所提出的文本分类方法，其安装在一个配备 Intel Core i5 CPU 和 Windows10 系统的个人电脑上。

为了构建用于实验的维吾尔文论坛文本集合，本文从人民网维文版、天山网论坛、ulinix 论坛和新疆公安数据库中收集约 2 400 篇论坛文本。其中，这些文本共分为 5 类，分别为正常类、暴恐类、反动类、色情类和迷信类，每种类型的文本数量不少于 200 篇。在均衡考虑各类样本比例下，将文本集的 60%作为训练样本集，其与 40%作为测试样本集。

3.2　性能度量

分类器的性能通常使用精确度（precision）和查全率（recall）来描述，精确度表示一个随机文本 d_x 被划分到第 c_i 类中，并且分类正确的概率。查全率表示随机文本 d_x 应当属于 c_i 类，并且这一决策被采纳的概率。精度和查全率表达式为

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \tag{11}$$

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \tag{12}$$

其中：TP_i 表示被正确分为第 i 类的文本数量；FP_i 表示被错误分到第 i 类的文本数量；FN_i 表示本属于第 i 类但被错误分到其他类的文本数量。

综合考虑准确性和查全率才能更好地表征分类器的性能。通常可以使用 F1 度量来对两个参数进行组合，其表达式为

$$F1_i = \frac{2Precision_i * Recall_i}{Precision_i + Recall_i} \tag{13}$$

3.3　词干提取性能分析

对于 *N-gram* 统计模型词干提取方法，其中需要设定 2 个参数，即 N 值和相似性阈值 T_S。N 值较大，提供了更多的语义信息，有助于提高分类器精度，但会大大增加特征项，提高计算复杂度；若 N 值较小，则产生的特征项所包含的语义信息也较少，区别性不强。

为了确定最优参数，分别设定 N=2、3、4 和 5，相似性阈值 T_S 分别设定为 0.6、0.7、0.8 和 0.9。构建 16 种参数组合，并在各种参数下进行词干提取和分类实验。为了公平比较，后续特征选择方法都采用 BMITS，分类器都采用 Rocchio 分类器。最终分类的 F1 度量值如图 2 所示。

图 2 的结果显示，随着 N 值的增加，分类器性能有所提高，但也会加大计算量。可以看到，当使用五字母组合(N=5)时的性能和四字母组合(N=4)时的性能相近，只有略微的提升。考虑到计算量，最终选择 N=4。另外，当相似度阈值 T_S =0.8 时，可以取得最好的分类效果；当 T_S =0.9 时，各种字母组合词干提取法的结果都变差。这是因为当阈值太高时，一些共享相同词根但相似度不够高的词将会被分开而不进行合并，所以降低了词干提取效果。最终，选择 N=4、T_S =0.8 作为词干提取方法的参数。

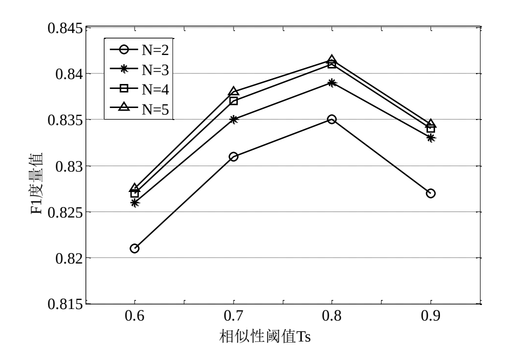


图 2　词干提取参数对分类性能的影响

表 2 展示了一组在 T_S =0.8 且 N=4 时 *N-gram* 统计模型提取的词干。

表 2　*N-gram* 统计模型提取的词干举例

反动类	پارچىلانماق (分裂)	ئاغدۇرۇش (推翻)	مۇستەقىل (独立)	خائىنلىق (叛变)	جاسۇس (间谍)
暴恐类	قاتىللىق (杀人)	بۇلاش (抢劫)	سوقۇشماق (打架)	مىلتىق (枪支)	زەخمەت (伤害)

3.4　术语选择性能分析

在对输入文本进行词干提取后，会形成一个具有大量术语的词集合，因此必须通过术语选择找到最具价值的术语子集。这里通过实验比较了本文 BMITS 选择方法与传统 MITS、χ^2 统计量、NGL 系数和 GSS 系数方法。其中，设置所选择的最终特征集(术语集)大小在 1000~5000 变化。为了公平比较，都采用相同的词干提取参数和 Rocchio 分类器。术语选择方法对分类性能的影响如图 3 所示。

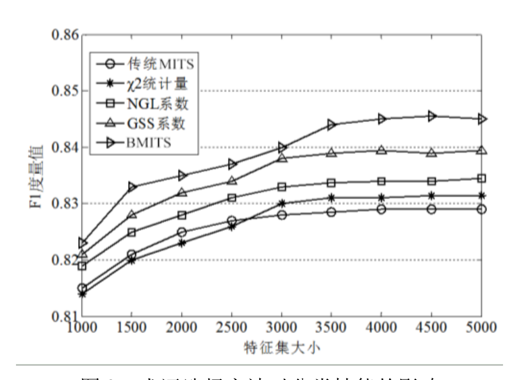


图 3　术语选择方法对分类性能的影响

图 3 中的实验结果表明，随着术语选择后的特征集大小的增加，各种方法的分类性能也会随之提高，但当特征集大小超过 3 500 时，性能趋于稳定，这就说明合适的特征集大小对分类器的性能有影响。特征集太小，不能包含所有有效术语；特征集太大，冗余术语数量也会增加，且加大了分类器的计算负载。为此，在本文术语选择步骤中，设置特征集大小为 3 500。

通过各种方法的比较看出，χ^2 统计量和传统 MITS 的性能较弱，相比而言，NGL 和 GSS 系数的表现较为良好。然而本文提出的 BMITS 方法获得了最佳性能。这是因为 BMITS 中引入

了候选术语与类别之间的互信息，均衡考虑了相关性和冗余性，所以能够选择出最佳术语集。

3.5 整体性能比较

为了评估本文方法的整体性能，将其与几种现有方法进行比较，如文献[5]提出的基于组合统计量(Dme)和 k-NN 分类器的维吾尔语分类方法(Dme+k-NN)、文献[6]提出的基于 TF-IDF 和 SVM 的方法(TF-IDF+SVM)、文献[7]提出的基于 N-gram 模型和 Manhattan 距离的方法(N-gram+Manhattan)。分类性能的比较结果如表 3 所示，同时给出了完成所有测试样本分类所需的时间。

表 3 可以看出，文献[5]方法所需时间最长，这是因为其通过融合 3 个度量来进行词干提取和术语选择，需要大量的计算。而文献[7]方法的时间较少，这是因为其没有采用例如 k-NN 等基于监督学习的分类器，只是通过一种距离度量来分类，计算较为简单。但是本文方法所消耗的时间最短，这是因为本文 Rocchio 分类器使用类别广义矢量代替了语料库中所有的训练文本矢量，在时间消耗上要明显优于 k-NN 和 SVM 分类器。此外，由 BMITS 术语选择过程产生了精简且有效的术语子集，有效提高了分类性能。

表 3　各种方法的分类性能

方法	精确度(%)	查全率(%)	F1 度量	时间(分钟)
文献[5]	81.47	82.73	82.09	6.8
文献[6]	82.66	80.95	81.79	5.7
文献[7]	82.68	83.87	83.27	4.2
本文方法	83.79	85.31	84.54	3.8

4　结束语

本文提出了一种应用在维吾尔文网页论坛文本过滤中的不良文本分类方法。通过预处理删除停止词，基于 N-gram 模型进行词干提取，通过 BMITS 进行术语降维，最后利用 Rocchio 进行文本分类及过滤。为了获得方法的各种最佳参数，通过大量实验来进行优化选择。最终的文本过滤实验结果证明了提出的方法能够用于维文论坛的信息过滤，具有有效性。

在未来研究中，希望通过进行一些更多的预处理工作来提高分类效果，例如从每个类别的局部术语库中选择术语，而不是从整个语料库的全局术语库中选择。另外，还可以研究其他分类器，如朴素贝叶斯以及神经网络分类器等，选择一种更加合适的分类器。

参考文献：

[1] 刘磊，李壮，张鑫，等．中文网络文本的语义信息处理研究综述 [J]. 计算机应用研究, 2015, 32 (1): 6-10, 16.

[2] 程俊霞，李芝棠，邹明光，等．基于 SVM 过滤的微博新闻话题检测方法 [J]. 通信学报, 2013, 34 (2): 74-78.

[3] 亚力青·阿里玛斯，哈力旦·阿布都热依木，陈洋．基于向量空间模型的维吾尔文文本过滤方法 [J]. 新疆大学学报：自然科学版, 2015, 32 (2): 221-226.

[4] Zhang B, Xu M, Wu M. Research on web filtering technology based on the dual feature selection [C]// Proc of IEEE International Conference on Network Infrastructure and Digital Content. Piscataway, NJ: IEEE Press, 2013: 675-679.

[5] 阿力木江•艾沙，吐尔根•依布拉音，艾山•吾买尔，等．基于机器学习的维吾尔语文本分类研究 [J]. 计算机工程与应用, 2012, 48 (5): 110-112.

[6] 热依莱木•帕尔哈提，孟祥涛，艾斯卡尔•艾木都拉．基于区分性关键词模型的维吾尔语本情感分类 [J]. 计算机工程, 2014, 40 (10): 132-136.

[7] 买买提依明•哈斯木，吾守尔•斯拉木，维尼拉•木沙江，等．基于 N 元模型的维吾尔语文本分类技术研究 [J]. 计算机应用研究, 2015, 32 (7): 1986-1988, 2004.

[8] Mi C, Yang Y, Wang L, et al. Detection of loan words in uyghur texts [J]. Communications in Computer & Information Science, 2014, 49 (6): 103-112.

[9] 阿不都萨拉木·达吾提，于斯音·于苏普，艾斯卡尔·艾木都拉．类别区分词与情感词典相结合的维吾尔文句子情感分类 [J]. 清华大学学报：自然科学版, 2017, 57 (2): 197-201.

[10] Froud H, Lachkar A, Ouatik S A. A comparative study of root-based and stem-based approaches for measuring the similarity between arabic words for arabic text mining applications [J]. Advanced Computing An International Journal, 2012, 3 (6): 12-19.

[11] Hadni M, Ouatik S A, Lachkar A. Effective arabic stemmer based hybrid approach for arabic text categorization [J]. International Journal of Data Mining & Knowledge Management Process, 2013, 3 (4): 1-14.

[12] 姜志威，丁晓青，彭良瑞，等．低数据资源条件下基于结构信息共享的无切分维文文档识别字符建模 [J]. 电子与信息学报, 2015, 37 (9): 2103-2109.

[13] Uchyigit G. Experimental evaluation of feature selection methods for text classification [C]// Proc of International Conference on Fuzzy Systems and Knowledge Discovery. Piscataway, NJ: IEEE Press, 2012: 1294-1298.

[14] Hoque N, Bhattacharyya D K, Kalita J K. MIFS-ND: a mutual information-based feature selection method [J]. Expert Systems with Applications, 2014, 41 (14): 6371-6385.

[15] Sowmya B J, Chetan, Srinivasa K G. Large scale multi-label text classification of a hierarchical dataset using Rocchio algorithm [C]// Proc of International Conference on Computation System and Information Technology for Sustainable Solutions. Piscataway, NJ: IEEE Press, 2016: 291-296.

[16] Selvi S T, Karthikeyan P, Vincent A, et al. Text categorization using Rocchio algorithm and random forest algorithm [C]// Proc of the th International Conference on Advanced Computing. Piscataway, NJ: IEEE Press, 2017: 124-129.